2В

УДК 004.931’1

Е.В. Волченко
Государственный университет информатики и искусственного интеллекта,
г. Донецк, Украина
LM@mail.promtele.com

Построение обучающей выборки w-объектов на основе коллективного решения группы экспертов

Работа посвящена решению задачи построения решающих правил в адаптивных системах распознавания при наличии классификации каждого объекта группой независимых экспертов. Для оценки степени согласованности экспертов в классификации объектов предлагается использовать показатель уверенности классификации. Для учета степени согласованности экспертов осуществляется переход к взвешенным выборкам w-объектов. Предлагается единый подход к формированию взвешенной выборки w-объектов по исходной выборке и добавляемым в процессе работы системы объектам. Анализ результатов тестовых исследований показал существенное снижение ошибок классификации при использовании выборки w-объектов для построения решающих правил классификации.

Введение

При построении обучающихся систем распознавания в большинстве случаев единственной априорной информацией, по которой выполняется построение решающих правил классификации, является обучающая выборка, содержащая данные о значениях признаков распознаваемых объектов и соответствующих этим объектам классах. Классификация объектов обучающей выборки в общем случае осуществляется экспертом и считается верной, поскольку проверить её правильность не представляется возможным [1]. При этом неверная классификация даже незначительного количества обучающих объектов может существенно изменить решающие правила классификации и привести к значительному ухудшению качества распознавания [2].

Для решения этой проблемы наиболее часто используется два подхода. В первом происходит отказ от имеющейся классификации объектов, выполняется кластеризация объектов обучающей выборки и по её результатам каждому объекту ставится в соответствие номер класса, полученный автоматически [2], [3]. Такой подход является единственно возможным, когда нет возможности получить дополнительную информацию о классификации объектов обучающей выборки. При этом очевидно, что отказ от имеющейся априорной информации может приводить к ухудшению качества распознавания [4]. Согласно второму подходу, используются данные о классификации объектов коллективом независимых экспертов, и классификация объектов определяется посредством обработки результатов частных классификаций этих экспертов [5], [6]. Если такие данные априорно не могут быть получены, то в качестве экспертов может выступать множество решающих правил, построенных по исходной выборке [2], [7].

Анализ многих прикладных задач, например, задач медицинской диагностики, для которых была известна классификация объектов группой экспертов, показал, что объекты, наиболее удаленные от межклассовой границы, относятся экспертами к одному из классов системы практически единогласно. Объекты, находящиеся в простран-

стве признаков вблизи межклассовой границы, достаточно часто получают неоднозначную классификацию экспертами. Использование в обучающих выборках только номера класса, которому отдало предпочтение большинство экспертов, и отсутствие учета степени согласованности классификации объектов экспертами, на наш взгляд, также может привести к ухудшению качества решающих правил классификации. Поэтому в работе [8] для оценки степени разногласия экспертов в классификации объектов обучающей выборки было предложено вычислять коэффициент уверенности классификации, являющийся, в то же время, критерием определения классификации объектов обучающей выборки.

Учет взаимного расположения объектов обучающей выборки является одним из наиболее эффективных способов повышения качества распознавания в обучающихся системах [9]. В наибольшей степени это проявляется для объектов, значения признаков которых изменяются динамически [10]. Изменение распознаваемых объектов с течением времени требует постоянного обновления обучающей выборки и, как следствие, корректировки решающих правил классификации. Системы, обеспечивающие такие возможности, получили название открытых адаптивных систем распознавания.

В работе [11] для учета расположения объектов в пространстве признаков была предложена идея перехода от традиционных обучающих выборок к взвешенным выборкам. Было показано, что использование взвешенных выборок помимо решения задачи сокращения обучающих выборок, являющейся одной из центральных задач построения открытых адаптивных систем распознавания, позволяет повышать эффективность систем за счет учета расположения объектов в многомерном пространстве признаков.

Данная работа является продолжением исследований в области построения открытых адаптивных систем распознавания и посвящена разработке единого подхода к построению взвешенных обучающих выборок на основе коэффициента уверенности классификации.

Постановка задачи

Пусть имеется некоторая конечная обучающая выборка объектов $X_i = \{X_1, X_2, ..., X_l\}$. Каждый объект X_i описывается системой признаков, т.е. $X_i = \{x_{i1}, x_{i2}, ..., x_{in}\}$, и представляется точкой в линейном пространстве признаков, т.е. $X_i \in R^n$. Для каждого объекта известна его классификация s экспертами $y_i = \{y_{i1}, y_{i2}, ..., y_{is}\}$, $y_{ij} \in V$, $V = \{V_1, V_2, \ldots, V_K\}$ – множество классов системы. Каждый эксперт характеризуется рейтингом R_j, $j = \overline{1, s}$.

Необходимо сформировать классифицированную взвешенную обучающую выборку w-объектов $X^W = \{X_1^W, X_2^W, \ldots, X_l^W\}$, $y_i^W \in V$, где $X_i^W = \{x_{i1}, x_{i2}, ..., x_{in}, p_i\}$, p_i – вес i-го w-объекта, y_i^W – классификация i-го w-объекта с учетом совокупного мнения о классификации объектов обучающей выборки всех экспертов.

Построение взвешенной обучающей выборки w-объектов по исходной выборке

Определение классификации объектов обучающей выборки при условии наличия множеств экспертных оценок для них предлагается выполнять путем расчета показателя уверенности классификации [8]. Его основу составляет рейтинг экспертов R_j,

оценивающий степень доверия классификации объектов, выполненной этим экспертом. Отметим, что если рейтинг экспертов неизвестен, то он может быть принят за единицу, т.е. $R_j = 1,\ j = \overline{1,s}$.

Определение. Показателем уверенности классификации $U(X_i / y_i = j)$ назовем отношение суммарного рейтинга экспертов, относящих объект обучающей выборки X_i к классу j, к общему рейтингу всех экспертов, т.е.

$$U(X_i / y_i = j) = \frac{\sum_{t=1}^{s}(R_t \cdot p_t)}{\sum_{t=1}^{s} R_t},$$

где $p_t = \begin{cases} 1, \text{если } X_i \text{ отнесен } t\text{-м экспертом к классу } j \\ 0, \text{иначе} \end{cases}$.

Определение классификации каждого из объектов обучающей выборки осуществляется путем выбора номера класса, соответствующего максимальному показателю уверенности классификации:

$$y_i = \arg\max_{j=\overline{1,k}} U(X_i / y_i = j).$$

По результатам определения классификации объекта X_i исходной обучающей выборки формируется w-объект X_i^W следующим образом:

1) признаки w-объекта X_i^W являются признаками объекта X_i исходной выборки;

2) объект X_i^W относится к классу, для которого $U(X_i / y_i = j)$ максимален;

3) вес p_i w-объекта X_i^W принимается равным максимальному значению показателя уверенности классификации (т.е. значению показателя уверенности классификации класса, к которому отнесен рассматриваемый объект).

В результате расчета показателя уверенности классификации по всем классам системы для всех объектов и определения максимальных из них, формируется взвешенная классифицированная обучающая выборка w-объектов $X^W = \{X_1^W, X_2^W, \dots, X_l^W\}$. Отметим, что в отличие от стандартного подхода в определении классификации объектов при наличии коллективной классификации группой экспертов, когда определяется только принадлежность объекта к одному из классов системы [12], предлагаемый подход позволяет оценить степень уверенности экспертов в правильности классификации и дает дополнительные исходные данные для дальнейшего построения решающих правил классификации.

Пополнение взвешенной обучающей выборки w-объектов

Одним из основных отличий открытых адаптивных систем распознавания является возможность добавления новых обучающих объектов на всем протяжении времени работы системы [10], что в свою очередь требует корректировки обучающей выборки и адаптации решающих правил классификации.

Возможны следующие ситуации при добавлении новых объектов:

1) если новый объект обучающей выборки классифицирован группой экспертов, то его обработка аналогична обработке объектов исходной выборки (рассчитывается показатель уверенности классификации и выполняется построение нового w-объекта);

2) если новый объект обучающей выборки классифицирован только одним экспертом, то значения признаков нового w-объекта приравниваются значениям признаков добавляемого объекта, вес нового w-объекта устанавливается равным единице и объект относится к классу, определенному единственным экспертом.

Определение рейтинга экспертов, представленных решающими правилами классификации

Как отмечалось ранее, получение классификации объектов обучающей выборки группой экспертов не всегда возможно. В этом случае в качестве экспертов используется множество решающих правил, построенных по исходной выборке [7]. Для расчета показателя уверенности классификации в таком случае необходимо определить рейтинг каждого из используемых решающих правил классификации. На наш взгляд, является естественным использовать в качестве рейтинга величину, характеризующую качество распознавания решающим правилом объектов тестовой выборки, т.е.

$$R_j = 1 - \frac{N(F(X'))}{|S(X')|}, \; j = \overline{1,q},$$

где $N(F(X'))$ – количество неверных классификаций объектов тестовой выборки X' решающим правилом $F(X')$;

$|S(X')|$ – размер тестовой выборки X';

q – количество используемых решающих правил классификации.

Полученный таким образом рейтинг экспертов (решающих правил) используется при расчете показателя уверенности классификации аналогично заданным рейтингам экспертов.

Классификация объектов по взвешенной обучающей выборке w-объектов

Основным отличием используемых алгоритмов построения решающих правил от множества известных алгоритмов является необходимость учета веса w-объектов. Так, для классификации распознаваемых объектов с использованием взвешенной обучающей выборки может быть использован модифицированный метод k-ближайших соседей. Классификация объектов определяется по k ближайшим w-объектам к классифицируемому объекту X'_i по следующей метрике:

$$F_{ij} = \frac{p_i \cdot p_j}{r_{ij}^2} = \frac{p_i \cdot p_j}{\left\| X'_i - X_j^W \right\|} = \frac{p_i \cdot p_j}{\sum_{o=1}^{n} \left(x_{io} - x_{jo} \right)^2}, \qquad (1)$$

где $p_i = 1$ – вес распознаваемого объекта, который принимается равным единице.

Два объекта являются ближайшими, если значение, рассчитанное по формуле (1), максимально.

Объект X'_i относится к тому классу, объектов которого среди k ближайших больше.

Результаты экспериментальных исследований

Для оценки эффективности применения предложенного в данной работе подхода был выполнен ряд экспериментальных исследований, в которых для обучающих выборок, классификация объектов которых выполнялась группой экспертов, сформированы выборки w-объектов и оценена эффективность построенных по ним решающих правил. Экспериментальные исследования проводились на исходных выборках размером 500 – 2000 объектов, значения признаков которых распределены по нормальному и равномерному законам распределения. Количество одинаковых классификаций объектов экспертами устанавливалось пропорционально удаленности объекта от межклассовой границы (наиболее удаленный объект был классифицирован всеми экспертами одинаково, лежащий на границе – с минимальным перевесом в сторону одного из классов).

Для оценки эффективности классификации использовались тестовые выборки размером 200 объектов, созданные по тем же генераторам, что и исходные обучающие выборки. В качестве критерия оценки эффективности использовалась частота неверных классификаций объектов тестовой выборки:

$$Z(X') = \frac{N(F(X'))}{|S(X')|},\ j = \overline{1,s}\ .$$

Результаты оценки являются средними по результатам 50 экспериментов.

Для оценки эффективности использования показателя уверенности классификации и взвешенных выборок, построенных на его основе, использовались следующие значения:

1) частота ошибочных классификаций $N(F_1(X'))$ объектов тестовой выборки по обучающей выборке, классификация объектов которой определялась большинством голосов экспертов [12];

2) частота ошибочных классификаций $N(F_2(X'))$ объектов тестовой выборки по обучающей выборке, классификация объектов которой определялась на основе показателя уверенности классификации;

3) частота ошибочных классификаций $N(F_3(X'))$ объектов тестовой выборки по взвешенной обучающей выборке w-объектов.

Результаты экспериментальных исследований при изменяющемся размере обучающих выборок и степени пересечения классов в пространстве признаков приведены в табл. 1 и 2 соответственно.

Анализ полученных результатов показывает, то использование показателя уверенности классификации для определения классификации объектов обучающих выборок позволяет в среднем уменьшить частоту неверных классификаций на 1,5%, а использование взвешенной выборки w-объектов – на 4,5%.

Отметим, что предложенный подход наиболее эффективен для обучающих выборок большого объема и классов, существенно пересекающихся в пространстве признаков, что является характерным особенностями адаптивных открытых систем распознавания.

Таблица 1 – Частота неверной классификации объектов тестовых выборок при изменяющемся размере обучающих выборок по 10% степени пересечения классов в пространстве признаков

Размер выборки	$N(F_1(X'))$	$N(F_2(X'))$	$N(F_3(X'))$
500	0,04	0,032	0,019
1000	0,033	0,029	0,017
1500	0,03	0,022	0,012
2000	0,028	0,019	0,006

Таблица 2 – Частота неверной классификации объектов тестовых выборок при различной степени пересечения классов в пространстве признаков при размере обучающих выборок 1000 объектов

Степень пересечения классов	$N(F_1(X'))$	$N(F_2(X'))$	$N(F_3(X'))$
0	0,008	0,006	0,001
10	0,033	0,029	0,017
20	0,074	0,057	0,031
30	0,119	0,105	0,076
40	0,208	0,185	0,139

Выводы

В работе предложен общий подход к построению взвешенных обучающих выборок w-объектов в открытых адаптивных системах распознавания по исходным обучающим выборкам и объектам, добавляемым в процессе работы систем при наличии данных о классификации объектов группой независимых экспертов. Для учета степени согласованности классификации экспертами предложено использовать показатель уверенности классификации, который во взвешенных обучающих выборках используется в качестве веса w-объектов. Проанализированы возможные виды добавляемых обучающих объектов и предложены способы построения по ним w-объектов. Описаны особенности классификации распознаваемых объектов на основе метода k-ближайших соседей по взвешенной обучающей выборке w-объектов. Результаты экспериментальных исследований по оценке эффективности использования взвешенных обучающих выборок w-объектов показали, что использование в качестве веса w-объектов показателя эффективности классификации позволяет в среднем на 4,5% уменьшить количество неверных классификаций. При этом наибольшее снижение частоты неверных классификаций наблюдается при существенном пересечении классов и обучающих выборках большого размера.

Литература

1. Лапко А.В. Непараметрические модели распознавания образов в условиях малых выборок / А.В. Лапко, В.А. Лапко, С.В. Ченцов // Автометрия. – 1999. – № 6. – С. 105-113.
2. Ширяев В.И. Использование адаптивных методов распознавания образов в задачах принятия решений / В.И. Ширяев, Б.М. Кувшинов // Искусственный интеллект. – 2002. – № 4. – С. 526-533.
3. Дуда Р. Распознавание образов и анализ сцен : [пер. с англ.] / Р. Дуда, П. Харт. – М. : Мир, 1976. – 510 с.
4. Загоруйко Н.Г. Прикладные методы анализа знаний и данных. – Новосибирск : Издательство института математики, 1999. – 270 с.
5. Городецкий В.И. Методы и алгоритмы коллективного распознавания: обзор / В.И. Городецкий, С.В. Серебряков // Труды СПИИРАН. – Вып. 3, Т. 1. – СПб. : Наука, 2006. – С. 139-181.
6. Файнзильберг Л.С. Обучаемая система поддержки принятия коллективного решения группы независимых экспертов / Л.С. Файнзильберг // Управляющие системы и машины. – 2003. – № 4. – С. 62-67.
7. Мазуров В.Д. Комитеты в задачах оптимизации и классификации / В.Д. Мазуров. – М. : Наука, 1990. – 248 с.
8. Волченко Е.В. Метод построения взвешенной обучающей выборки при групповой экспертной классификации // Двенадцатая национальная конференция по искусственному интеллекту с международным участием КИИ-2010: Труды конференции. – М. : Физматлит, 2010. – Т. 2. – С. 90-97.

9. DuMouchel B. Squashing flat files flatter / B. DuMouchel, C. Volinsky, T. Johnson, C. Cortes, D. Pregibon // Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining. – New Jersey, AAAI Press, 1999. – P. 6-15.
10. Pal S.K. Pattern Recognition Algorithms for Data Mining : Scalability, Knowledge Discovery and Soft Granular Computing / S.K. Pal, P. Mitra. – Chapman and Hall/CRC, 2004. – 280 p.
11. Волченко Е.В. Метод построения взвешенных обучающих выборок в открытых системах распознавания / Е.В. Волченко // Доклады 14-й Всероссийской конференции «Математические методы распознавания образов (ММРО-14)», (Суздаль, 2009). – М. : Макс-Пресс, 2009. – С. 100-104.
12. Миркин Б.Г. Проблема группового выбора / Б.Г. Миркин. – М. : Наука, 1974. – 256 с.

О.В. Волченко

Побудова навчальної вибірки w-об'єктів на основі колективного рішення групи експертів

Роботу присвячено дослідженню задачі побудови вирішуючих правил в адаптивних системах розпізнавання за наявності класифікації кожного об'єкту групою незалежних експертів. Для оцінки міри узгодженості експертів в класифікації об'єктів пропонується використовувати показник упевненості класифікації. Для врахування міри узгодженості експертів здійснюється перехід до зважених вибірок w-об'єктів. Пропонується єдиний підхід до формування зваженої вибірки w-об'єктів по вихідній вибірці і об'єктам, що додаються в процесі роботи системи. Аналіз результатів тестових досліджень показав істотне зниження помилок класифікації при використанні вибірки w-об'єктів для побудови вирішуючих правил класифікації.

E.V. Volchenko

Construction of the W-Objects Training Sample on Basic of Set of Experts' Solution

A work is devoted to solving the problem of constructing decision rules in adaptive recognitions systems in the presence of classification of each object by the group of independent experts. To estimate consistency of experts in the objects classification it is proposed to use the index of classification's confidence. Transitions to the weighted samples of w-objects are made to take into account consistency of experts in the classification of objects. The unified approach to the formation of the weighted sample of w-objects from the original sample and adding of the objects are proposed. Significant reductions in classification errors when using the sample of w-objects in the construction of decision rules of classification are shown.

Статья поступила в редакцию 09.08.2010.